IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 003, LIR098
DOI: https://doi.org/10.52731/lir.v003.098

Analysis of College Students' Career Awareness after Taking First-Year Career Courses Using the Structural Topic Model

Tatsuya Tsumagari [*], Yoko Nakazato [†], Takashi Tsumagari [‡]

Abstract

This study analyzed career awareness as a result of first-year career education courses using the Structural Topic Model (STM). Career awareness refers to topics mentioned in students' free-writing reports on career understanding. We examined career awareness from two perspectives. First, we examined the career awareness of students who were learning the target subjects without any influence from year-to-year fluctuations. Second, we focused on topics consistently generated as students' career awareness and examined the differences in career awareness of each student type classified according to their interest in the lecture content, focusing on the difference in the word distribution of the topics. To examine these two perspectives, we extracted topics through STM analysis to set year and student type as covariates of topic prevalence, and student type as a covariate of topic content. The results showed that there were three topics of career awareness that remained stable regardless of year: "self-strengths and weaknesses," "encounters with others with different values," and "toward life fulfillment." We found that for "toward a fulfilling life," one type of students tended to use words with a long time span meaning, while another type of students tended to use words with a short time span meaning.

Keywords: actual state of learning, career education, first-year experience, learning evaluation, structural topic model

1 Introduction

What do students who take first-year university career courses learn about their careers? Many studies measure course effectiveness by comparing pre- and post-course surveys. However, this does not always reveal the reality of students' learning. To solve this problem, a textual analysis of students' free writing was attempted to evaluate learning [1].

One text analysis method is the topic model [2]. A topic model is a method that explores multiple subjects (topics) that appear in a document and the words that are closely related to those topics to help understand the document. Topic models infer topics based on textual data. They do not allow for correlations between topics, nor do they allow for setting covariates. For this reason, it is common to evaluate the relationship between covariates and topics ex-post after estimating a topic model if one wants to examine covariates. However, incorporating information related to the corpus structure into the topic model, modifying the prior distribution, and partially pooling information among similar documents is superior to ex-post evaluation [3].

[*] Seigakuin University, Saitama, Japan
[†] Kagoshima University, Kagoshima, Japan
[‡] Prefectural University of Kumamoto, Kumamoto, Japan

This analysis is made possible by the Structural Topic Model (STM). STM allows for the presence of covariates related to topic prevalence and content, and is an extension of the topic model [3]. It also allows for correlations between topics. It has been reported that the STM is superior to the topic model for analyzing the relationship between covariates and topics [4].

This study reports the results of an analysis using the STM to understand the actual state of learning in first-year career education. We used two covariates, year and student type, which were classified based on lecture interest. Two analyses were conducted for each of the covariates; one was to estimate the topics that were stably generated, regardless of the year. This analysis allowed us to understand the career awareness that students learn stably in first-year career education courses. The second was an analysis of the differences in career awareness based on students' interest in the lecture content. Even when students attend the same lecture, they learn differently. This implies that even if we talk about the same topic, students may perceive it differently. Using student type as a covariate, it was possible to identify the effects of different student types on the same topic.

The above report shows that the analysis of students' free writing reports by STM is a useful method for evaluating learning.

2 Method

2.1 Sample

The sample comprised first-year students who took a first-year career course at a public university in Japan from 2017 to 2022. This required first-year course is offered from April to August. The total number of all students taking it was 1,780 from 2017 to 2022. After completing all the lectures, the students submitted two assignments. One assignment was to select two lectures that they were interested in, and the other was a free writing report of approximately 600 characters (in Japanese) describing "how they understood the concept of careers" throughout the lectures. We used the free-writing reports of 1,732 students, excluding 16 students with missing data from the 1,748 students who volunteered to participate in the survey. The reports were anonymized prior to the analysis. The average number of characters in the freewriting reports was 622 (standard deviation=67).

2.2 Method of Analysis

2.2.1 Classifying students by their interests

The first-year career course sample comprised seven lectures. These lectures were classified into three types based on content: A) theory type (four lectures), B) self-understanding type (one lecture), and C) role model type (two lectures).

The theory type is a lecture that explains the meaning of learning at university and the significance of career development for the future, while the self-understanding type is a lecture that encourages self-reflection on current abilities using the results of a test evaluating generic skills. The last lecture, the role-model type, provided stories about senior students' and graduates' experiences as familiar role models for first-year students. Students were asked to choose two of the seven lectures they found interesting. Based on the combination of these choices, the students were classified into five types.

Table 1 lists the number and composition of each type. When all lectures were selected with equal probability, the expectation values of the ratios were 28.6% for Type I, 19.0% for

Type II, 38.1% for Type III, 9.5% for Type IV, and 4.8% for Type V. Compared with the expectation values, Type I was low, and types IV and V were high. This result shows that of the students surveyed were strongly interested in self-understanding and role models, while they were less interested in theoretical-type lectures that included many relatively long-term career topics.

Table 1: Student types (n=1,732)

Type of Student	Combination of lectures that the students were interested in	Number of students	Ratio (%)
I	A: Theory, A: Theory	97	5.6%
II	A: Theory, B: Self-understanding	226	13.0%
III	A: Theory, C: Role model	477	27.5%
IV	B: Self-understanding, C: Role model	612	35.3%
V	C: Role model, C: Role model	320	18.5%

2.2.2 Estimation of Structural Topic Model

This study examined actual state of learning using STM to analyze students' free-writing reports on their understanding of their careers. The year and student type were set as covariates related to topic prevalence, which is a component of STM, and student type was set as a covariate related to topic content. Assuming changes specific to a particular year, a cubic spline function was specified that allowed nonlinear changes to be confirmed.

The R package stm [5] was used as the analysis tool, and Mecab [6] was used for morphological analysis of the free-writing reports.

1) Preparation of Data set

We analyzed 1,732 free-writing reports. First, a morphological analysis of the data from 1,732 students was conducted, and we extracted nouns (general and proper nouns), verbs (independent), adjectives, adverbs, and unknown words. Using these results and freewriting reports as references, we preprocessed the text data for dictionary registration, stop words, and synonyms. For dictionary registration, a list of 299 words was registered as a user dictionary to handle words that could not be processed by Mecab, such as unique expressions. The stop words selected were the words that include numerical values such as "one" and "two" (8 words), and words such as "do" and "think" (12 words) that are difficult to interpret as topic content among frequently appearing words. As for letter variation, the target words were identified as follows: full-size and half-size, case quirks (e.g., PROG TEST, prog test), notational quirks caused by kanji conversion, notational quirks caused by synonyms (e.g., advantages and strengths), and speaker quirks (e.g., senior student, person's name), and each was assigned to a unique word. For speakers, "special lecturers (chairman, president, dean of faculty, and lecturers within and outside university)," "senior students," and "graduates" were assigned as unique words. In some cases, the words "special lecturer," "senior students," and "graduates" appeared simultaneously in one sentence in the free response reports, and the description of that sentence was a reference to all speakers. To facilitate interpretation of the analysis results, when the words "special lecturer," "senior student," and "graduates" appeared in one sentence, they were grouped together and converted to the word "speaker."

We extracted nouns (general and proper nouns), verbs (independent), adjectives, adverbs, and unknown words that appeared in the reports of more than five students from the pre-processed data. Following this process, 1, 510 words were obtained. We created a dataset from these 1,510 words including the data representing the number of occurrences of each word in reports and metadata of year and student type of reports and estimated the STM. All analyses were conducted in Japanese, and the final notations were converted into English.

2) Considering of number of topics

When estimating STM, the number of topics must be determined a priori. However, there is no established method for determining the number of topics. In this study, held-out likelihood and residual dispersion were used as evaluation indices to determine the number of topics. We calculated and compared these two evaluation indices based on the number of topics ranging from 5 to 30. The held-out likelihood ranged from -6.15 to -6.00, and the residual dispersion showed a decreasing value starting from topic number 9 (Figure 1). Based on the above, nine topics (held-out likelihood = -6.05, residual dispersion = 2.05) were judged appropriate.

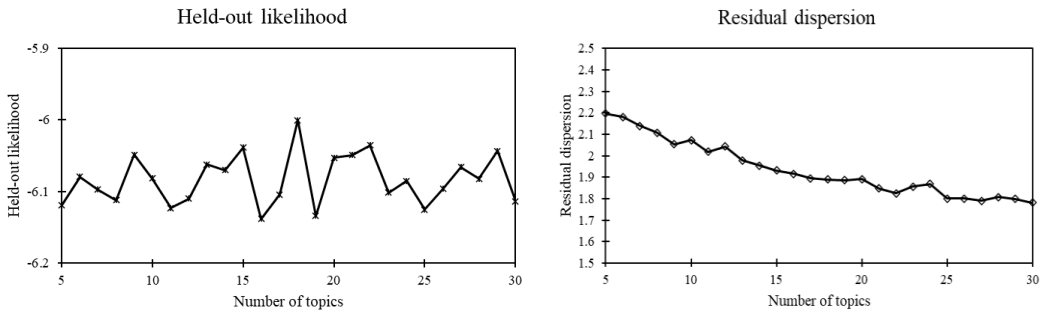

Figure 1: Evaluation index for considering of number of topics

3) Topic Labeling

Topics were labeled by careful consideration of the authors, referring to the content of the reports that had the highest percentage of the most frequently occurring top words the corresponding topic contained in each of the four indicators: Highest Prob, FREX, Lift and Score. According to Ishida[7], “Intuitively, Highest Prob is a group of words estimated to have the highest probability of occurring for each topic, FREX is a group of words that characterize the topic, Lift is a group of words that are particularly likely to appear for the topic, and Score is an indicator similar to the TF-IDF for frequency information, and is the top ranked word group for which all topic distributions are taken into account.” [7]. According to Bischof et al. [8], exclusive words are suitable topic summary words. Based on these considerations, this study mainly referred to FREX word groups in labeling to make sense of the topics.

3 Results and Discussion

3.1 The Results of Extracting Topics

Table 2 shows the topic proportions, results of the most frequently occurring words in the FREX index, and labels assigned to the nine topics extracted from the free-writing reports on career understanding.

Referring to the nine topics, students described what a career is in their free-writing reports(e.g., " Career is a journey of life" and "Career development learning from role model"), as well as the awareness toward future career development (e.g., "self-strengths and weaknesses," "encountering others with different values," "finding dreams and goals," etc.). The most common topics were in the area of awareness of future career development. This is a satisfactory result for the first-year experience of students, as it encourages them to become aware of their career development.

Table 2: Proportion of topics extracted, top FREX frequency words and labels

Topic	Proportion	FREX：Frequently occurring words	Label
1	0.116	Disadvantage, advantage, develop, PROG TEST, competency literacy, low, task, ability, know	Self-strength, weakness
2	0.092	People, exist, encounter, communication, thought, relationships, make, other people, value, different	Encounter others with different values
3	0.126	Progress, life, speaker, life, words, role, self, talk, achieve, reflect	Career is a journey of life
4	0.154	Do, dream, goal, find, interest, decide, field of vision, set, make effort, know	Find dreams and goals
5	0.153	Build up, brush up, knowledge, gain, work, skill, qualification, process, ability	Attitude toward work
6	0.069	Way, learning, learn, discipline, alive, place, faculty, civil servant, intelligence, build up	University as a learning place
7	0.147	Senior students, graduates, very, leave, impression, hear, especially, special lecturer, talk, story about career development experience	Career development learning from role model
8	0.113	Build up, life, spend, university student, might, high school, lead, future, nothing	Toward life fulfillment
9	0.029	Make up, to be able, broaden, others, people, process, self, human, being worth doing, improve	Self-development through relationship with others

3.2 Year Dependency of Topics

The syllabus of the course was the same each year, and there was no year dependency for the course content. However, the lecturers change annually. In addition, the area where the surveyed university is located suffered significant earthquake damage in 2016, and the disaster experience caused significant changes in students' career awareness [9]. It is possible that this effect continued to a small extent in subsequent years. In 2020-2022, the COVID-19 pandemic caused all lectures to change from face-to-face to online. These external factors may have changed the topic distribution from year to year.

Figure 2 shows the year-to-year changes in topics. Three topics, topics 1, 2, and 8, showed no year dependence at the 5% significance level. These three topics, "Self-strengths and weaknesses," necessity for "Encountering others with different values" and "Toward life fulfillment," were not affected by external factors in the targeted first-year career courses and were consistently learned in all years.

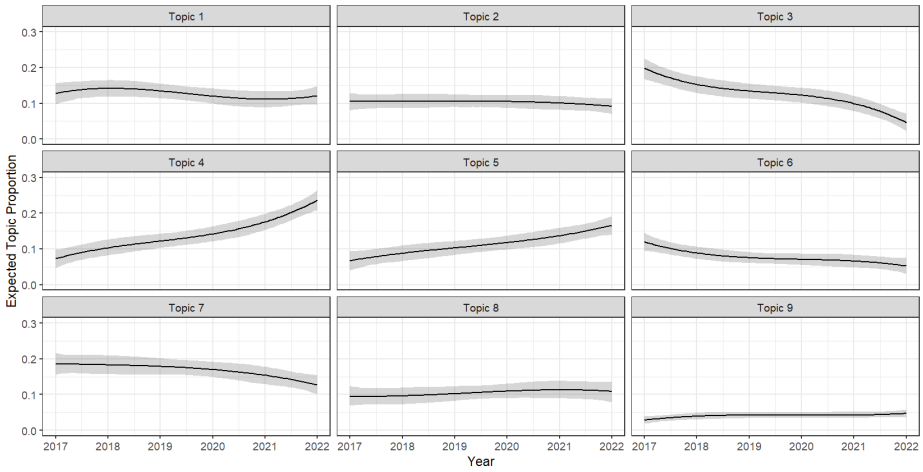

Figure 2: Year change for each topic (shading indicates 95% confidence interval)

3.3 Narratives by Student Type on the Same Topic

We examined the differences between student types in topics 1, 2, and 8, which were generated as stable student career awareness. A nonlinear regression model was used to analyze each topic as a dependent variable, with year and student type as explanatory variables for the covariates of topic proportions. The reference category for the student type was set as Type V (role model inclination). The results showed that the topic proportion did not differ by student type for most topics. However, there was a statistically significant difference in the topic proportion of topic 1 for Type II students compared to Type V students ($p=0.002$), but the difference was small in the aggregate. We can say that there was almost no difference in topic proportions among different types of students.

However, it is possible that different student types use different words to discuss the same topic. Table 3 shows the results of extracting the most frequently appearing words in the FREX index by student type for topics 1, 2, and 8. The most frequently appearing words for topic 1 showed no significant differences among the student types, and it seems that all types of students used the same words to describe their strengths and weaknesses in topic 1.However, for topic 2, it is characteristic that the word "quality" appears at the top of the list for student type I, which is not at the top for the other types. This result might argue that type I students talk about encounters not only with others, but also with a certain level of "quality" taken into account. Furthermore, the word "friends" appears for Type V students, and students of this type described about their encounters with friends at the university. Finally, for topic 8, Student type I used words such as "proud, female, perceive, life, project, life plan, correct, way of thinking," which are words that have a long-time range, such as life as a woman or life plan. However, the top frequently appeared words for student types III, IV, and V are "high school," "university," and "summer vacation," suggesting that they tend to view the topic of "life fulfillment" within the short range of the present moment as a student. The topic "Finding dreams and goals" (topic 4) contained the word "job hunting" as one of the frequent appeared words especially for student type V, which is not covered by any of the other types of students. This indicates that type V students tended to focus on the near future. The results for topic 4 also confirmed the characteristics of Student Type V in topic 8.

Table 3: Top Frequent Words in Indicator FREX by Student Type on Topic 1, 2, and 8

Topic	Student type	FREX: Frequently occurring words
1	I	Advantage, PROG test, improve, disadvantage, ability, low, competency, communication, task, information
	II	Disadvantage, advantage, task, PROG test, level, ability, competency, low, high, information
	III	Ability, PROG test, disadvantage, advantage low, improve, task, literacy, competency, information
	IV	Competency, literacy, improve, disadvantage, basic, PROG test, task, advantage, low inter-personal
	V	PROG test, advantage, competency, information, disadvantage, low, literacy, task, know, part
2	I	Person, quality, thought, exist, other people, tool, different, relate, attitude
	II	Communication, person, personal connection, way, exist, thought, represent, people, aspect, leadership
	III	Person, change, bad, thought, exist, communication, make, worth, other people, different
	IV	Encounter, person, make, relationship, encounter, exist, worth, way of thinking, other people, high
	V	Person, encounter, thought, communication, relationship, exist, friend, different, relate, worth
8	I	Proud, female, perceive, life, project, life plan, plan, correct, way of thinking, spend
	II	Life, enjoy, spend, self, decide, live, nothing, lead, how, go
	III	High school, spend, children, life, go, change, long, choice, plan, lead
	IV	Build up, university student, ideal, life, high school student, choice, goal, school, children, limit
	V	Future, might, book, summer vacation, spend, nothing, life, read, lead, how

4 Conclusion

This study examined the career awareness that students learned after taking a career development course by analyzing 1732 first-year students' free-writing reports about their career understanding of first-year career education courses from 2017 to 2022. We used a Structural Topic Model for the analysis, and nine topics were extracted for students' career awareness.

The results showed that three of the 9 topics ("Self-Strengths and Weaknesses," "Encounters with Others with Different Values," and "Toward a Fulfilling Life") had stable topic proportions regardless of the year of learning. Although the syllabus for the courses was the same for all the years analyzed, different lecturers were on stage in different years, and the lecture style varied greatly due to the Coronavirus Disease 2019 Pandemic. The results suggest that these three topics were not affected by differences in the syllabus from year to year and were stable topics in the career awareness of the students who took the course.

Different types of Students may use different words to discuss the same topic. Therefore, focusing on the word distribution of topics, we analyzed the differences in career awareness among five student types classified according to their interests. Examining the above three topics, which were stable regardless of the year, we confirmed the differences among student types on the topics of "encountering others with different values" and "toward life fulfillment." For example, on the topic of "Toward life fulfillment," Type I students tended to view and describe the topic in a long time range, while Type V students tended to view and describe about the topic in a short time range.

This study identified topics related to career awareness in first-year career courses that were stable and independent of the year using STM. In addition, we found that different types of students have different perspectives on stable topics, and we believe that an evaluation method

using STM that utilizes students' freewriting reports may be useful for learning evaluation.

This study was limited to topics that were stable and independent of the year, and other topics could not be examined. The analysis and discussion of these topics will be an issue for future research.

Acknowledgement

This work was supported by JSPS KAKENHI Grant No. 21K02634.

References

[1] M. Kikuchi, T. Suda, Y. Tange and K. Murakami, “Students’ Learning and Thought-Inducing Factors Analyzed from Their Comments on a Career Course” [in Japanese], Jpn. Assoc. for College and University Education, Vo.41, No.1,2019, pp.117-156.

[2] T.L. Griffiths and M. Steyvers, “Finding Scientific Topics”, Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl 1), 5228-35, 2004.

[3] M. E. Roberts, B. M. Stewart, D. Tingley, and E. Airoldi, “The structural topic model and applied social science.” Neural Information Processing Society,2013, Copy at http://www.tinyurl.com/y67skpul.

[4] M. E. Roberts, B. M. Stewart, and E. Airoldi, “A Model of Text for Experimentation in the Social Sciences”, Journal of the American Statistical Association, 111, 988-1003, 2016.

[5] M. E. Roberts, B. M. Stewart, and D. Tingley, “stm: An R Package for Structural Topic Models”, J. Stat. Soft., vol. 91, No. 2, 2019, pp.1-40.

[6] https://taku910.github.io/mecab/

[7] M. Ishida, “Text Mining by R” [in Japanese], Morikita-Shupan, 2020, pp.23-55.

[8] J. M. Bischof and E. M. Airoldi, “Summarizing topical content with word frequency and exclusivity”, In Proceedings of the 29th International Conference on Machine Learning (ICML-12),2012.

[9] Y. Nakazato, Tatsuya Tsumagari, Takashi Tsumagari, “Analyzing the characteristics of first-year university students' career awareness through free-writing reports”, Proc. 9th International Congress on Advanced Applied Informatics, IIAI-AAI 2020, 2020, pp.347-350.